Expressive-Oriented Time-Scale Adjustment for Misplayed Musical Signals Based on Tempo Curve Estimation

03/Sept/2013 @ DAFx 2013

Graduate School of Computer and Information Sciences, Hosei University, Japan

Yuma KOIZUMI, Katunobu ITOU
Deviances in musical performances

- Relevance between **deviances** and **performance expression**

**Deviances in music**

<table>
<thead>
<tr>
<th>Tempo (timing)</th>
<th>Timbre</th>
<th>Pitch</th>
<th>Amplitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>accelerando, ritardando</td>
<td>Sul tasto, Sul ponticello</td>
<td>vibrato, tremolo, portamento</td>
<td>crescendo, legato, marcato</td>
</tr>
</tbody>
</table>

- Low-proficiency performances include other type of deviance; **mis played components** ⇒ due to poor control of instruments

**Adjustment:**

Estimate intended tempo fluctuation and remove mis-played components
True tempo curve & mis played components

- Sophisticated players control the tempo smoothly ⇒ tempo curve
- Amateur players cannot control the tempo smoothly

Observed tempo = true tempo curve + mis played components
**True tempo curve & mis played components**

- Sophisticated players control the tempo smoothly ⇒ **tempo curve**
- Amateur players cannot control the tempo smoothly

---

**Our goal**

\[
\text{Observed tempo} = \text{true tempo curve} + \text{mis played components}
\]

Estimate

Remove
Tempo estimation using onset timing

\[ \text{note duration } [n] \text{ (sec)} = \frac{60 \text{ (sec/min)}}{\text{tempo } [n] \text{ (beats/min)}} \times \text{note value } [n] \text{ (beats)} \]

onset timing [n] (sec) = \( \sum_{m=1}^{n-1} \) note duration [m]

\[ y[n] = \sum_{m=1}^{n-1} \frac{60}{b[m]} h[m] + e[n] \]

- \( y[n] \) = onset timing
- \( b[n] \) = true tempo curve
- \( h[n] \) = note value
- \( e[n] \) = mis played component

※ Note value:
- Crotchet note = 1.0
- Quaver note = 0.5
Flowchart of the proposed method

Input data
- Audio signal
- Note numbers and note values

Onset detection
- Onset candidate set
- $F_0$ contour
- Score alignment
- Onset selection

Tempo curve estimation and audio adjustment

Output data
- Adjusted audio signal
Onset detection (1/2); Complex Mel KLD

- An accurate onset detection method is required
- An appropriate acoustic feature depends on the type of instrument and the playing style
- In any performance, audience can recognize onset timing
- Modeling of an acoustic feature for aural characteristics:

complex mel-spectrum KL divergence (CMKLD)

\[
D[k] = \sum_{\psi} \left| S_{\psi,k} \log \frac{S_{\psi,k}}{\hat{S}_{\psi,k}} \right|
\]

\[
= \sum_{\psi} \left| S_{\psi,k} \right| \sqrt{\left( \log \left( \frac{S_{\psi,k}}{\hat{S}_{\psi,k}} \right) \right)^2 + \left( \phi_{\psi,k} - \hat{\phi}_{\psi,k} \right)^2}
\]

- Guitar: Amplitude spectrum
- Violin: Phase spectrum
Onset detection (1/2); Complex Mel KLD

\[
D[k] = \sum_{\psi} \left| S_{\psi,k} \log \frac{S_{\psi,k}}{\hat{S}_{\psi,k}} \right|
\]

Weighting coefficient for harmonic structure

Create onset candidates by calculating CMKLD and peak picking

\[
\begin{align*}
S_{\psi,k} & : \text{observed complex mel-spectrum} \\
\hat{S}_{\psi,k} & : \text{predicted complex mel-spectrum} \\
\phi_{\psi,k} & : \text{observed phase mel-spectrum} \\
\hat{\phi}_{\psi,k} & : \text{predicted phase mel-spectrum} \\
\psi & : \text{Mel-log frequency}
\end{align*}
\]
Onset detection (2/2); onset selection

Score information

<table>
<thead>
<tr>
<th>n</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Note value</td>
<td>0.5</td>
<td>0.5</td>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>Note number</td>
<td>67</td>
<td>69</td>
<td>69</td>
<td>71</td>
</tr>
</tbody>
</table>

Observed F0

CMKLD & onset candidates

Score alignment

Onset selection
Flowchart of the proposed method

Input data
- Audio signal
- Note numbers and note values

Onset candidate set
- \( F_0 \) contour

Onset selection

Tempo-curve estimation & Signal adjustment

Tempo curve estimation and audio adjustment

Output data
- Adjusted audio signal
True tempo curve estimation (1/2)

\[ y[n] = \sum_{m=1}^{n-1} \frac{60}{b[m]} h[m] + e[n] \]

- \( y[n] \): onset timing
- \( b[n] \): true tempo curve
- \( h[n] \): note value
- \( e[n] \): mis-played component

**Polynomial tempo curve model**

\[ b[n]^{-1} = \sum_{p=0}^{P} w_p \left( \sum_{m=1}^{n} h[m] \right)^p \]

Note duration

\[ y[n+1] - y[n] = \Delta y[n] = 60 \left( \sum_{p=0}^{P} w_p \left( \sum_{m=1}^{n} h[m] \right)^p \right) h[n] + \Delta e[n] \]

where \( \Delta e[n] = e[n+1] - e[n] \)
True tempo curve estimation (2/2)

\[ \Delta y[n] = 60 \left( \sum_{p=0}^{P} w_p \left( \sum_{m=1}^{n} h[m] \right)^p \right) h[n] + \Delta e[n] \]

Assume: mis played component \( e[n] \) has a Gaussian distribution [Joder, et al., 2011]
\[ \rightarrow \Delta e[n] \] also has a Gaussian distribution

✓ Calculate the regression coefficients \( w_p \)
✓ Calculate true tempo curve

✓ Which is the optimal polynomial degree \( P \)?

- \( P = 2 \):
  - AIC = -93.2844
- \( P = 5 \):
  - AIC = -91.8715
- \( P = 8 \):
  - AIC = -88.9285

No. 1: \( \rightarrow \) determined by minimization of AIC (Akaike information criterion)
Signal adjustment

- Expansion and contraction of the step size of an inverse Fourier transform of the power spectrogram

By stretching the step size in each note, the tempo can be modified

Stretch factor \([n]\) = intended note duration \([n]\) \(\div\) observed note duration \([n]\)

\[
= \left(\frac{60}{b[n]} h[n]\right) \div (y[n + 1] - y[n])
\]

The stretch factors are multiplied by the original step sizes

\(\Rightarrow\) The adjusted signal is synthesized by inverse Fourier transform
Estimate true tempo curve, and adjust this recording.
Subjective evaluation

- Amateur musicians tried to mimic the professional performance
  \[\Rightarrow\text{Adjusted sound should resemble professional recordings}\]

- Subjects evaluated closeness of tempo fluctuation to the professional performance

![Graph showing MOS improvement and significant differences](image)

- MOS was improved & Significant differences were observed
Subjective evaluation (discussion)

MOS: In term of “impairment”

- 3 = “slightly annoying”
- 4 = “perceptible, but not annoying”

This “impairment” as bring due to poor control of instruments...

- 3 = slightly annoying your mis playing
- 4 = perceptible, but it didn’t matter to listen to your performance

The proposed method can...

- ✓ estimate the intended tempo fluctuation
- ✓ adjust the recording based on the true tempo curve
Conclusions

A time-scale adjustment method which estimates the intended tempo fluctuation and removes misplayed components was proposed.

Key techniques

- Intended tempo fluctuation: true tempo curve was estimated by polynomial regression analysis
- Time-scale adjustment: Expansion & contraction of power spectrogram

Closeness of rhythm,

MOS values have improved (3 → 4) for all instruments.

The proposed method can adjust misplayed recordings based on true tempo curves
Future work

- **Intra-note segmentation**
  - Note duration should be stretched only in the steady state

- **Micro-tempo fluctuation**
  - True tempo curve: Macro-tempo fluctuation
  - Intended slight tempo fluctuation should be also estimated
Q & A
Akaike Information Criterion

A measure of the relative quality of a statistical model, for a given set of data.

\[ AIC = 2P - 2 \ln(L) \]

\[ P = \text{the number of parameter in the model} \]
\[ L = \text{the likelihood function for the model} \]

The preferred model has minimum AIC value.

AIC not only rewards goodness of fit \((L)\), but also includes a penalty of the number of parameters \((P)\).

AIC of the proposed tempo curve estimation method.

\[ \sigma^2 = \frac{1}{N} \sum_N \left( \Delta y[n] - 60 \sum_P w_p g[n]^p h[n] \right)^2 : \text{Mean square error} \]

\[ AIC = N \ln(2\pi\sigma^2) + N + 2(P + 2) \]
Mel-scale (one of the logarithmic scales of frequency)

- A perceptual frequency scale of pitches
- The mel scale…
  - expresses aural perception better than linear frequency scale.
  - is often used in speech recognition (MFCC).
Evaluation; onset detection

- Mean absolute error between detected onset times and the manually selected ones.

- The experiments were performed on six violin solo phrases (in total 152 onsets), that include some expression marks (e.g. legato, feroce (wildly)).

Result

Mean absolute error was **18 ms**

- Less than 7% of a quarter note duration in the tempo of Allegro (BPM ≈ 120)
- Pitch perception requires 20-30ms

This accuracy can be considered to be adequate.